ADAPTIVE SENSING AND DESIGN FOR RELIABILITY

P. Singh, D. Sylvester, D. Blaauw

University of Michigan, Ann Arbor

email: {prsingh, sylvester, blaauw}@umich.edu

ABSTRACT

Chip lifetime degradation due to oxide break down is a major concern for today's designers. We review existing methods to solve the gate oxide reliability issues and also introduce an *in situ* degradation monitoring technique. This technique allows early detection of oxide degradation and makes a system aware of its reliability. When used in conjunction with reliability management schemes, it minimizes existing pessimistic reliability margins and allows an improvement in device performance.

Keywords: reliability, oxide degradation, sensors

INTRODUCTION

Technology scaling has resulted high levels of integration which has enabled sustained growth of the electronics industry. Advanced process nodes meet the performance requirements of today's electronic systems. However, it is becoming increasingly difficult to meet chip lifetime specifications while maintaining high yield. Modern technology scaling has abandoned the constant field paradigm, resulting in steady increases in gate oxide electric field in the past few generations. High electric fields increase the rate of degradation of the gate oxide, making gate oxide degradation a major bottleneck in sustaining the functionality of a chip throughout its lifetime [1]. This makes it harder to meet the reliability targets in advanced process nodes.

Designers have traditionally used static reliability management to meet reliability specifications. This method limits the supply voltage to a fixed maximum value for all fabricated chips such that a chip lifetime of the required level is attained. The lifetime of a chip, however, is a statistical variable due to the innate randomness in the degradation process. Moreover, process variation, fluctuations in environmental conditions such as voltage and temperature, and state dependence of the oxide degradation also add to the randomness in lifetime [2]. Hence the voltage limit is set so that the *weakest* chip meets the lifetime requirements under worst-case conditions. Hence, many chips will fail much later than the desired lifetime. The pessimism of static reliability management creates reliability slack on a chip-by-chip basis that can be traded off with performance to provide a significant performance benefit in cutting edge technology nodes [3].

To enable this tradeoff, a system needs to be aware of its reliability status. Such a system is then capable of dynamically adjusting the operating conditions of the chip (in particular, supply voltage and the maximum temperature limit) to achieve peak performance benefits while just meeting the lifetime specification. This approach has been referred to as Dynamic Reliability Management (DRM) [4, 5, 6]. These systems use on-chip sensors to get information that can be used to compute or predict the reliability state of the chip. DRM can be implemented in three ways: model based, degradation sensor based, and *in situ* monitoring based, as proposed in this paper.

In the degradation model-based approach, on-chip voltage and temperature sensors are employed to sense the operating conditions of the chip. The data from these sensors is fed to a degradation model which is used to compute the expected reliability state of the chip [4]. DRM systems based on this approach must account for inaccuracies in the sensors and the degradation models themselves by adding margins that make this approach more conservative. In addition, this approach does not address any innate sources of variation in lifetime such as those due to process variation, state-dependence and the inherent randomness of oxide breakdown, and hence has to be used with considerable margins.

The degradation sensor-based approach obviates the need for voltage and temperature sensors as well as degradation models. Instead, it relies on special sensors that directly monitor the degradation of replicated transistor oxides. Degradation sensors are distributed across the core in large numbers so that the sensors experience the same environmental conditions as the devices in the actual circuit. Degradation data from these sensors is then used to estimate the degradation of the actual devices in the chip [2]. Since such degradation sensors experience the same process conditions as the actual devices and also do not incur model inaccuracies, they can operate with tighter margins. However, degradation sensors do not account for lifetime variations due to innate randomness of degradation mechanisms, process mismatch between the replicated oxides that are monitored and the functional oxides of devices in the chip, and state dependence of stress. Thus, considerable margins remain.

The *in situ* monitoring-based DRM scheme employs a direct approach to monitor degradation. In this methodology the actual devices used in the circuit are measured directly to determine their degradation. This approach addresses all sources of lifetime variation and hence provides the most accurate observation of degradation, resulting in almost complete margin elimination. The key challenge is to achieve this with minimal invasiveness and overhead.

This paper reviews the static reliability management technique and the DRM schemes based on degradation models and degradation sensors. We also discuss a proposed *in situ* monitoring technique that allows the DRM system to be more aggressive with voltage scaling and hence minimizes reliability margins.

STATIC RELIABILITY MANAGEMENT

Breakdown of a MOSFET's gate oxide is a statistical process due to the innate randomness of the degradation mechanism. Moreover, factors including device process variation, fluctuations in operating conditions (voltage and temperature), and state dependence of stress further add to the randomness in device degradation. As a result, process engineers have to consider all these factors at design time before imposing a supply and temperature limit on the operating conditions of the devices. Fig. 1 shows the simulated lifetime distribution of an ensemble of chips under different process, voltage, and temperature conditions.

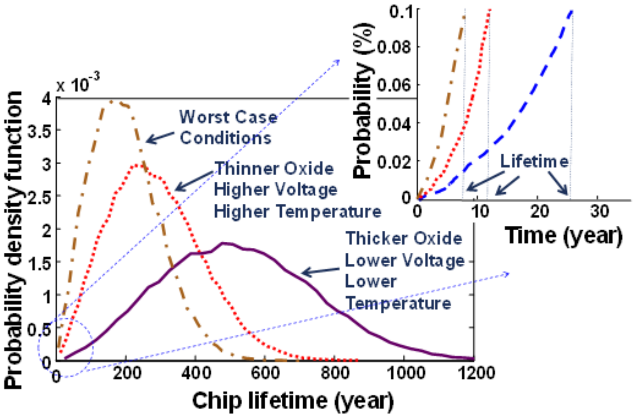

Fig. 1 Lifetime distribution of chips under different process, voltage and temperature (PVT) conditions. The desired lifetime of 10 years is not met under worst PVT conditions.

A percolation-based oxide degradation model was used in this simulation [7] . The figure shows that as the voltage and temperature are lowered and the oxide thickness reduced, the mean and spread in the lifetime of the chips decrease [3]. To ensure that a very low number of the chips fail (the desired yield) during the in-field operation of the chips, worst process, voltage, and temperature conditions are assumed to arrive at a lifetime distribution, which further reduces the mean lifetime of the chips. In reality these conditions will not be experienced by all chips at all times, making such assumptions very conservative.

If a lifetime of ten years and a yield of 99.9% is desired, then not more than 0.1% of the chips should fail at the end of ten years. As shown in Fig. 1 this does not hold true under the conservative worst process and operating condition assumptions. Hence, the voltage would have to be scaled down until the above mentioned criterion is met. This results in reduced performance of the chips and a lifetime much greater than the specification for many chips, or a *reliability slack,* owing to pessimistic assumptions. This slack can potentially be traded with performance enhancement by increasing the supply voltage. However, this requires that the system is self-aware of its reliability.

These systems can use DRM schemes to dynamically manage the reliability and the performance of the system. To employ DRM schemes the chip must be able to compute its reliability state. This can be accomplished by feeding temperature and voltage readings from sensors into a reliability model. However, accurate calibration of such a model with every technology node is non-trivial and resource intensive. Hence, in this paper we focus on DRM schemes that use reliability sensors as well as a proposed scheme that uses *in situ* reliability monitoring. We discuss each in the following two sections.

SENSOR-BASED RELIABILITY MONITORING

To employ DRM schemes, the chip must be able to compute its reliability state. The shortcomings of the model-based approach are partially overcome by degradation sensor-based reliability management. The idea is to use degradation sensors that consist of test devices exposed to the same voltage and temperature conditions as the core devices. Hence, these sensors are placed close to the core circuits to be monitored. However, the variation in degradation due to innate randomness cannot be captured by only a couple sensors; hundreds or even thousands of sensors are required to capture the statistics of degradation [2]. This imposes tight constraints on the area and the power consumption of these sensors. Moreover, it is important that these sensors track degradation very early after the onset of degradation so that the DRM scheme has enough time and flexibility to manage the chip reliability and react in time to changes in the degradation state.

Oxide degradation sensors have been proposed in [8, 9]. Fig. 2 shows the schematic of an oxide degradation sensor proposed in [8].

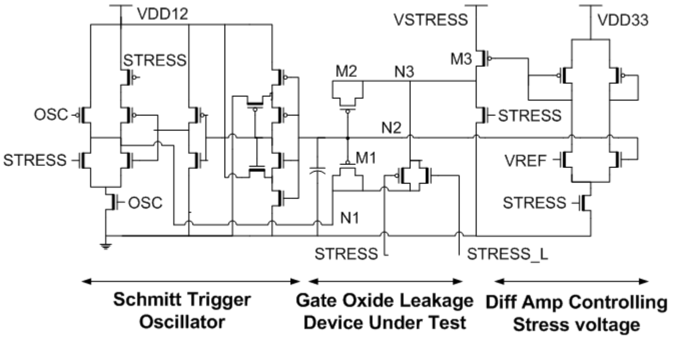
VDD12
STRESS
OSC
STRESS
OSC
VSTRESS
M2
N3
M3
VDD33
N2
M1
N1
STRESS
VREF
STRESS
STRESS
STRESS_L
Schmitt Trigger Oscillator
Gate Oxide Leakage Device Under Test
Diff Amp Controlling Stress voltage

Fig. 2 Gate-oxide degradation Sensor schematic.

The sensor consists of test devices M1 and M2, which are exposed by the stress voltage Vstress, which is the regular supply voltage of the chip or can be a higher voltage for accelerated testing. When a measurement is taken, the senor outputs the frequency of a ring oscillator, the delay of which is determined by the gate leakage of M1 and M2. Hence the frequency output of this sensor is directly proportional to the gate leakage of M1 and M2. As the gate oxide of M1 and M2 degrades, their gate leakage and the frequency output of the sensor rise. The digital output of the sensor makes it easy to process its data. The sensor has the area of 21 NAND3 gates, making it amenable for use in large numbers.

Fig. 3 shows degradation data collected using the sensor in [8]. The degradation captured by the sensors is gradual, allowing the DRM scheme to manage the operating conditions and control the degradation rate of the chip. For a small sample of 16 sensors, the gate leakage degradation varies from 5-40% over a die. The sensor data is used to collect statistics of degradation, which are then used to compute an upper and lower bound on the degradation. In a chip consisting of millions of transistors, this variation is much larger and hence there is a considerable difference between the upper and lower bounds of degradation. To avoid unexpected failures, the DRM scheme assumes the upper bound on degradation for all devices and adds some additional margin to this to determine the acceptable maximum supply voltage setting.

IN SITU RELIABILITY MONITORING

The sensor-based DRM scheme addresses the environmental conditions (voltage/temperature) and global process conditions among the sources of lifetime variation. To account for the other factors, such as local process variations and the innate randomness of the degradation, pessimism is introduced while computing the reliability state of the chip. This pessimism can be eliminated by the *in situ* monitoring of the degradation of the actual circuits as opposed to predicting their degradation using degradation sensors that lie alongside the circuit. This technique removes uncertainty due to local process, voltage, temperature conditions, the state dependence of degradation, as well as the innate randomness of the degradation process.

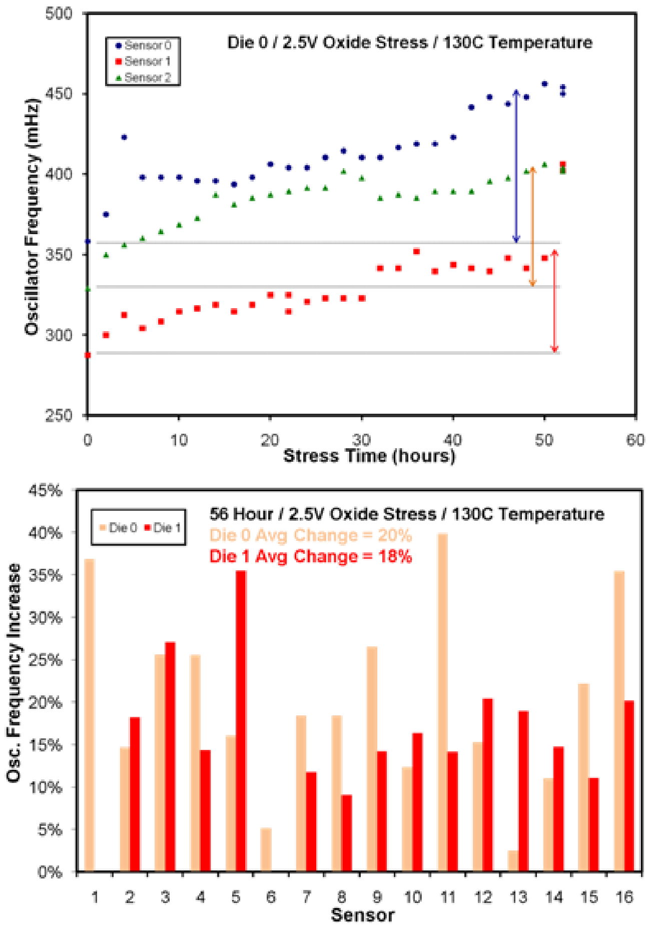

Fig. 3a (Top) Gradual gate-oxide degradation captured by the degradation sensor. Fig. 3b (Bottom) The gate current degradation varies from 5% to 40%.

In situ monitoring can be implemented in a number of ways. Delay has been proposed as a metric for degradation sensing; however, it is difficult to detect small changes in gate delay. In addition, the delay is sensitive to environmental conditions. Due to these factors a delay sensor's threshold to flag the onset of degradation must be set high to prevent any false alarms. This in turn would delay the degradation detection which may not leave enough time to take corrective action before the chip fails. Moreover, our silicon measurements show that a gate's delay might actually decrease with degradation, making the delay sensor-based *in situ* monitoring unreliable at times.

Gate leakage is the most direct measure of gate oxide degradation. However, it is difficult to measure the gate current in the presence of background currents such as subthreshold leakage, band-to-band tunneling, and gate-induced drain leakage. In addition, voltage and temperature strongly impact the total measured current, further complicating the measurement of changes in gate current.

In our proposed approach the key in detecting oxide degradation is sensing the change in the I_g-V_{gs} characteristics of a degraded device. In [10] the leakage of a degraded gate oxide is modeled as follows:

$$I_g = K\,(V_{gs})^P, \qquad (1)$$

where I_g is gate current and V_{gs} is gate-source voltage.

Fig. 4 illustrates how the values of K and P vary with degradation in time and consequently how the gate leakage increases. The values of K and P are based on measurements reported in [10].

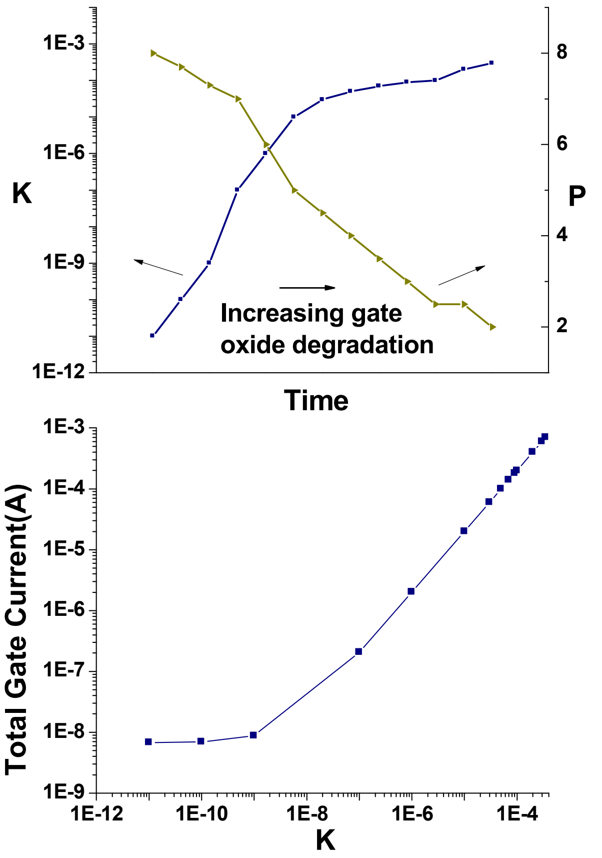

Fig. 4a (Top) Increasing gate-oxide degradation can be modeled as an increase in the value of K and decrease in the value of P. Fig. 4b (Bottom) The gate current increases by orders of magnitude due to gate-oxide degradation.

As the gate oxide degrades, defects, or trap sites, are formed in the oxide. The defects are at lower energy levels than the barrier height introduced by the insulator and hence the defects alter the exponential dependence of the gate current on voltage across the gate oxide.

Due to this phenomenon the I_g-V_{gs} characteristics of the device become more linear as a device degrades, which is illustrated by progressively straighter lines in Fig. 5. It is this key behavior that we use to detect the degradation of a device. We monitor this behavior for a cluster of gates to reduce monitoring overhead.

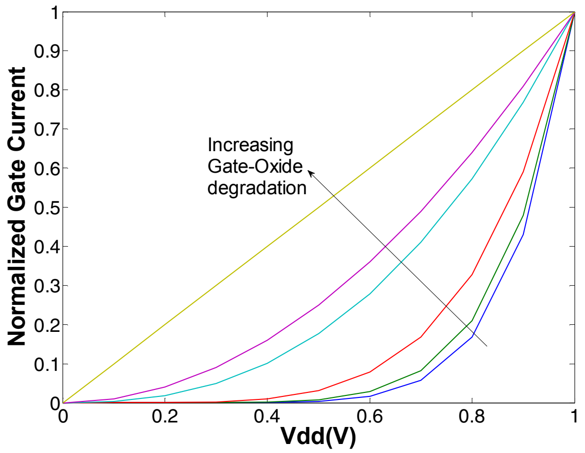

Fig. 5 The non linear nature of the Ig-Vgs characteristics of the gate-oxide becomes more linear with degradation.

As introduced in [11], the proposed *in situ* monitoring implementation leverages the prevalence of MTCMOS-based designs with PMOS header switches, a common technique to reduce standby power [12] with relatively low overhead.

Fig. 6 shows a circuit block partitioned into clusters, each connected to the power supply through a standard high-Vt MTCMOS switch and a weak PMOS device (WP) with a controllable gate voltage, Vbias.

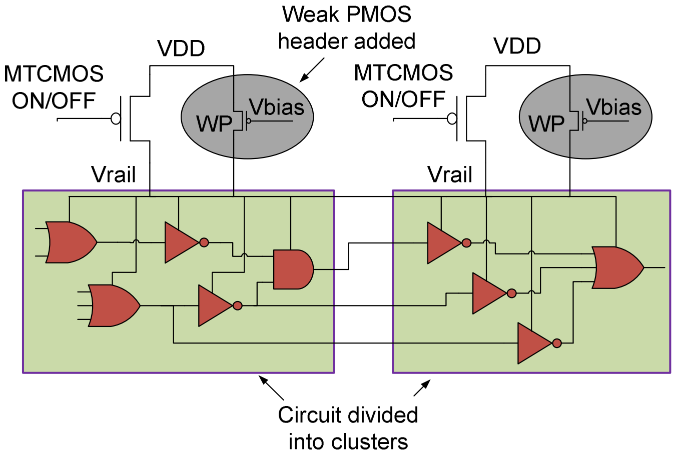

Fig. 6 The *in situ* monitoring technique is implemented by dividing a circuit into clusters using MTCMOS headers. Weak PMOS headers are added to monitor the conductance of the clusters.

The design is periodically taken offline and tested for oxide degradation by sweeping the gate voltage of WP from 0 to VDD using an on-chip DAC, while the virtual rail voltage is recorded using an on-chip ADC. The resulting Vbias vs. Vrail (V/V) curve is then analyzed to detect the onset of oxide breakdown (OBD). Initially, both oxide leakage and sub-threshold leakage are strongly non-linear with supply voltage, resulting in a characteristic "hockey stick" curve. However, as the device degrades the V/V curve flattens (Fig. 7).

Based on this key behavior shift, we define a figure of merit called the degradation voltage angle (DVA) that measures the angle of a straight line fitted to the V/V curve over the 90 – 10% Vrail interval. As a gate degrades, the oxide displays more linear resistive behavior and a sharp drop in DVA is observed (Fig. 8).

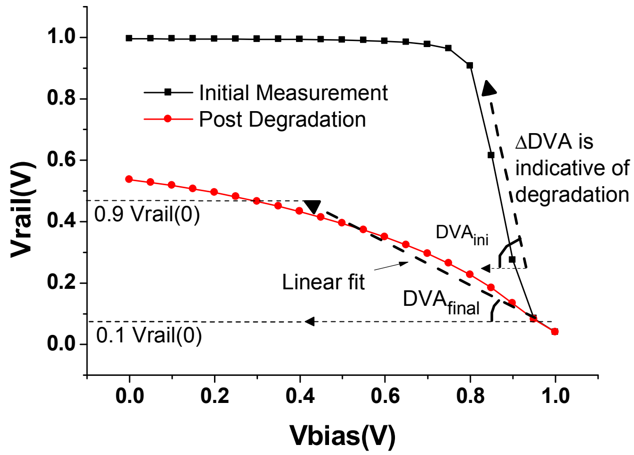

Fig. 7 The nature of Vrail vs. Vbias curve changes with degradation. DVA is defined to quantify this change in behavior.

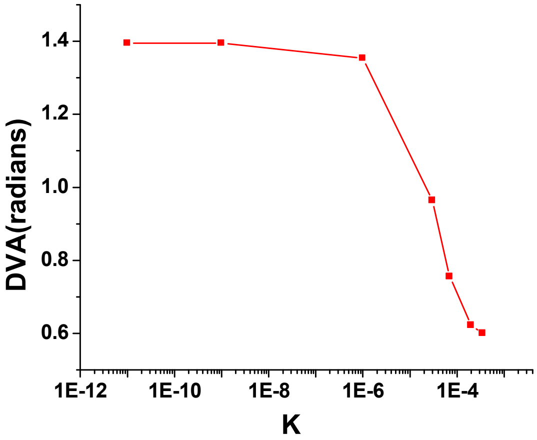

Fig. 8 DVA drops sharply with degradation which flags the onset of breakdown.

The technique was implemented in two test chips fabricated in 65nm CMOS. The nominal supply voltage in this 65nm technology is 1V. The first chip applies the technique to individual gates (INVERTER, NAND, NOR) and XOR parity trees (gate count ranging from 64-1024 gates) for a detailed study of the OBD effect. The stress voltage of 4V at a temperature of 125C is used for all the experiments on this test chip. The second chip implements a FIR filter to demonstrate applicability to larger circuit blocks. The technique can be applied to larger designs, in which case the overhead is further amortized. The stress voltage of 3V at a temperature of 165C is used for all the experiments on this test chip.

Fig. 9 shows the measured V/V, DVA, and gate delay curves for an inverter. We define 15% drop in DVA as the detection point of OBD onset. In this case the proposed technique detects the onset of gate-oxide degradation with as little as a 3% increase in the delay of an individual gate.

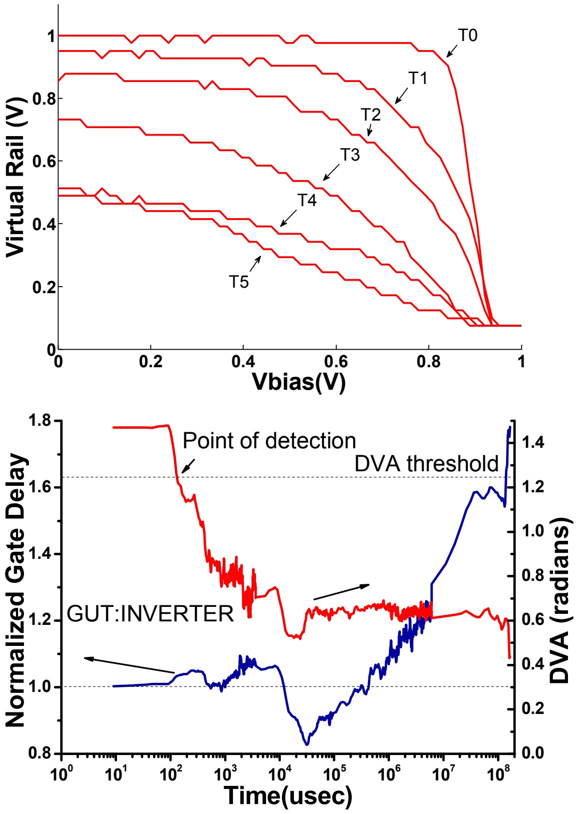

Fig. 9a (Top) Silicon measurement of Vrail vs Vbias curve for a stressed INVERTER at different points of degradation.
Fig. 9b (Bottom) Silicon measurement of DVA and Delay of a stressed INVERTER.

This illustrates that delay monitors like the one proposed in [13, 14, 15] will not be able to detect degradation until the delay is severely affected. Impact of stress on delay shows non-monotone behavior, at times resulting in faster gate delays. This is expected and is caused by the suppression of voltage swing under certain failure modes [10]. The corresponding degradation is captured by the DVA. Eventually the gate delay increases to 20X and then fails completely.

Since the subthreshold current is a strong function of temperature, the leakage of the non-failing gates can overwhelm the change in the gate leakage of the failing gate. To determine the sensitivity of the V/V curve to temperature, Fig. 10 shows the V/V, DVA, and path delay curves for an XOR tree consisting of 64 gates at 25C and 125C. There is a 10X change in subthreshold leakage for this 100C change in temperature, which changes the V/V curves significantly. However, the DVA metric changes by only 7% (less than the picked failure detection threshold of 15%), showing that DVA provides a robust measure of gate oxide degradation across temperature. The excellent robustness of DVA metric eliminates the need to calibrate or compensate for temperature, which would increase the complexity of the approach.

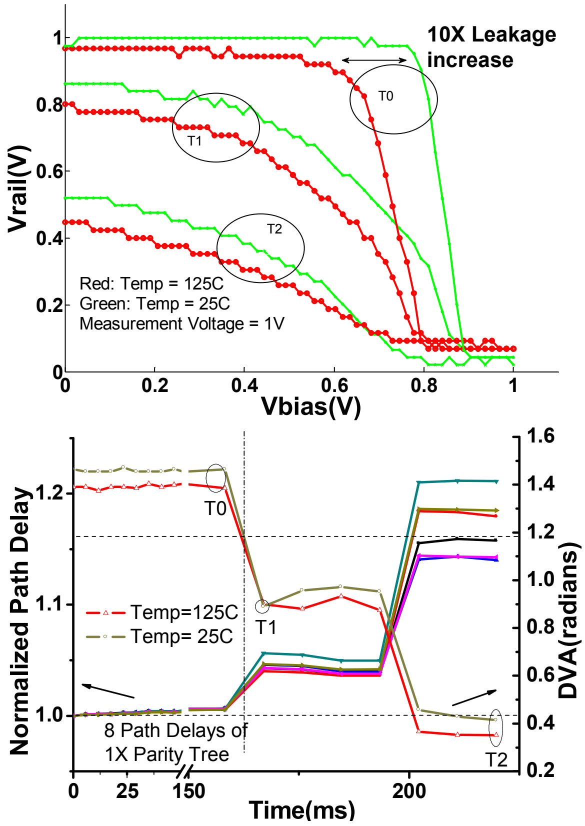

Fig. 10a (Top) Silicon measurement of Vrail vs Vbias curves of a 64 gate XOR parity tree at 25C and 125C at three different points in degradation showing immunity of DVA measure to environmental conditions.
Fig. 10b (Bottom) Measured DVA (at 25C and 125C) and Delay of a stressed 64 gate XOR parity tree.

Fig. 11 shows the effectiveness of this technique as the cluster size varies.

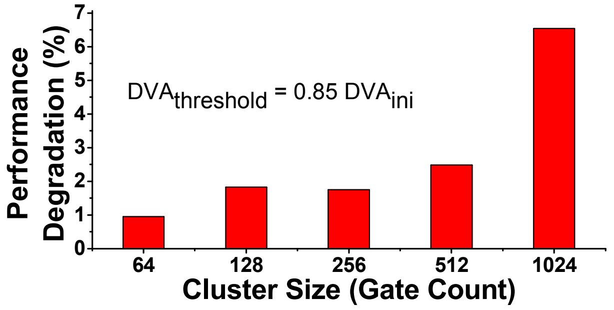

Fig. 11 Measurements show that the time to detection of onset of degradation increases with cluster sizes larger than 512 gates.

As the cluster size increases past 512 gates the failure detection is delayed since the leakage increase of the failing gate(s) is masked by the background leakage of the non-failing gates.

Fig. 12 shows a large variation in time to onset of OBD for 63 inverters, illustrating the statistical nature of oxide degradation.

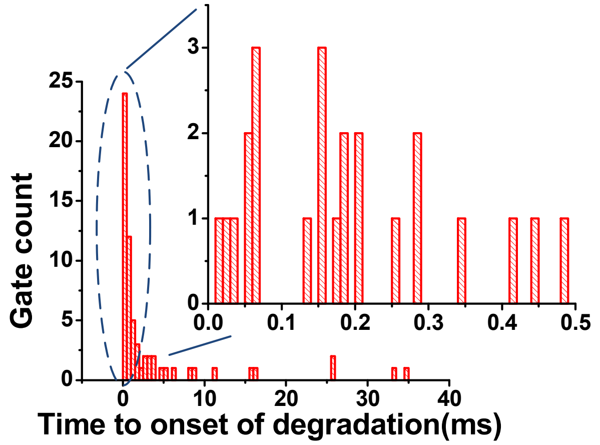

Fig. 12 Measurements show a large variation in the time to onset of gate-oxide degradation.

The second test chip applied the approach to a 16-bit, 8-tap FIR filter consisting of approximately 7K gates (Fig. 13).

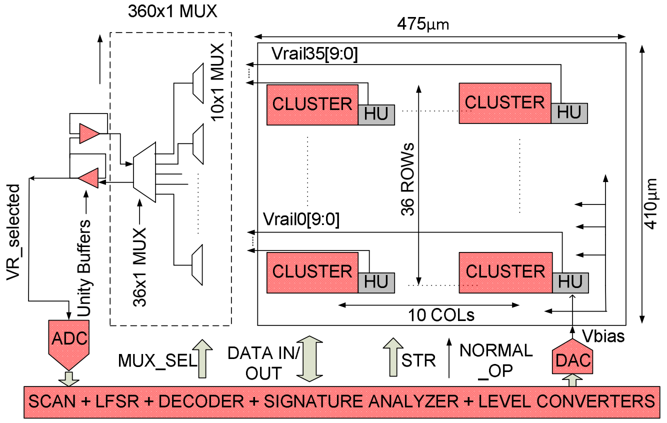

Fig. 13 *In situ* monitoring enabled FIR Filter architecture.

The FIR is divided into 360 clusters of ~20 gates placed into 36 rows and 10 columns using an automated design flow. To monitor each of the 360 virtual rails (VRs), a low leakage 360x1 two-stage mux is used. Since the VRs are driven by small leakage currents, it is extremely important to isolate the selected VR. To this end, a unity gain buffer mirrors the voltage seen on the selected VR onto the other non-selected VRs. The design area overhead of implementing this technique is 17% compared to a design without MTCMOS and 5% compared to a standard MTCMOS design. The overhead can be reduced by increasing the cluster size, which reduces the number of header devices and VRs.

Fig. 14 shows the clusters flagged for onset of degradation over time. Out of 360 clusters, 141 clusters were stressed and monitored. The first detected cluster failure corresponds to a performance degradation of the FIR by 0.5%.

All the silicon measurement results presented show that the *in situ* technique successfully monitors the degradation state of the actual devices in the chip and detects the onset of OBD very early in the chip's life. This provides the DRM controller with sufficient time to manage the chip's reliability, eliminating nearly all reliability slack and allowing the maximum performance from the whole ensemble of the chips, irrespective of PVT conditions.

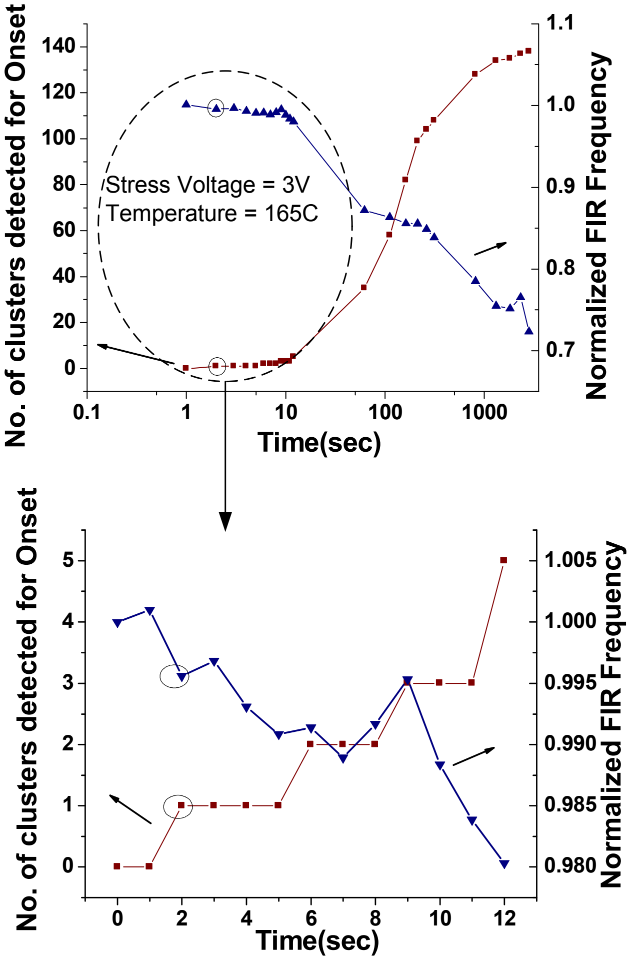

Fig. 14a (Top) The performance of the FIR degrades as clusters are detected with onset of gate-oxide degradation.
Fig. 14b (Bottom) The performance degradation is 0.5% when the first cluster is flagged for onset of degradation.

CONCLUSION

With technology scaling, the feature size of devices has reduced considerably while their supply voltage has not decreased proportionally. This yields performance gains at the cost of reduced device reliability. Gate oxide degradation is a major reliability threat to devices in advanced process nodes. To meet chip lifetime specifications, designers traditionally employ a "static reliability management" technique that does not address the sources of lifetime variations: 1) innate randomness in the degradation process, 2) process variation in the devices, 3) fluctuations in environmental conditions, and 4) circuit state. Hence, this approach tends to be very pessimistic and wastes considerable reliability slack that can be traded to recover performance losses.

Reliability-aware systems enable margin reduction by using different DRM schemes that are based on degradation models, degradation sensors, or *in situ* monitoring of the degradation of the actual core devices on the chip. Among these three methods, only the *in situ* monitoring approach addresses all the sources of lifetime variation and gives the most accurate degradation measurement.

We proposed a method for *in situ* monitoring of the gate oxide degradation of the *actual* devices of the core. The method enables early detection of the gate oxide degradation, providing sufficient time for a DRM controller to manage chip reliability while maximizing the performance gain from the whole ensemble of chips, irrespective of the sources of lifetime variations.

ACKNOWLEDGEMENTS

The authors would like to thank NSF and GSRC for funding support and ST Microelectronics for fabrication support.

REFERENCES

[1] J. H. Stathis, "Gate Oxide Reliability for Nano-Scale CMOS," in *International Conference on Microelectronics*, 2006, pp. 78-83.

[2] E. Karl, D. Blaauw, and D. Sylvester, "Analysis of System-Level Reliability Factors and Implications on Real-Time Monitoring Methods for Oxide Breakdown Device Failures," in *Proceedings of IEEE* International Symposium on Quality Electronic Design *(ISQED),*2008, pp. 391-395.

[3] Cheng Zhuo, David Blaauw, and Dennis Sylvester, "Post-Fabrication Measurement-Driven Oxide Breakdown Reliability Prediction and Management," in *Proceedings of IEEE/ACM International Conference on Computer-Aided Design,* San Jose, November 2009, pp.441-448.

[4] E. Karl, D. Blaauw, D. Sylvester, T. Mudge," Multi-Mechanism Reliability Modeling and Management in Dynamic Systems," in *IEEE Transactions On VLSI Systems*, April 2008, pp. 476-487.

[5] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers, "The case for lifetime reliability-aware microprocessors," in *Proceedings of 31st Annual International Symposium on Computer Architecture*, 2004, pp. 276–287.

[6] Z. Lu, W. Huang, M. R. Stan, K. Skadron, and J. Lach, "Interconnect lifetime prediction under dynamic stress for reliability-aware design," in *Proceedings of IEEE/ACM International Conference Computer-Aided Design*, 2004, pp. 327–334.

[7] R. Degraeve, et. al, "A consistent model for intrinsic breakdown in ultra-thin oxides," *International Electron Devices Meeting*, Dec. 1995, pp. 863-866.

[8] E. Karl, P. Singh, D. Blaauw, D. Sylvester, "Compact In-Situ Sensors for Monitoring Negative-Bias-Temperature-Instability Effect and Oxide Degradation," in *IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers*, 2008, pp.410-411.

[9] J. Keane, S. Venkatraman, P. Butzen, C. H. Kim, "An array-based test circuit for fully automated gate dielectric breakdown characterization," in *IEEE Custom Integrated Circuits Conference (CICC)*,Sept. 2008, pp. 121-124.

[10] R. Rodriguez, J. H. Stathis, and B. P. Linder, "Modeling and experimental verification of the effect of gate oxide breakdown on CMOS inverters," in *Proceedings of IEEE International Reliability Physics Symposium*, Dallas, TX, 2003, pp. 11–16.

[11] P. Singh *et al.*," Early Detection of Oxide Breakdown Through In Situ Degradation Sensing", in *IEEE International Solid State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2010, pp. 190–191.

[12] G. Gerosa *et al.*, "A sub 1 W to 2 W low power IA processor for mobile internet devices and ultra-mobile PCs in 45 nm high-k metal gate CMOS," in *IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 256–257.

[13] A. Drake *et al.*, "A distributed critical-path timing monitor for a 65 nm high performance microprocessor," in *IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 398–399.

[14] J. A. Blome, S. Feng, S. Gupta, and S. Mahlke, "Online timing analysis for wearout detection," in *2nd Workshop on Architectural Reliability(WAR-2)*, Dec 2006.

[15] T.W. Chen, K. Kim, Y. Kim and S. Mitra, "Gate-Oxide Early Life Failure Prediction," in *IEEE VLSI Test Symposium*, 2008.